A Practical Procedure for the Construction and Reliability Analysis of Fixed Length Tests with Random Drawn Test Items

Draaijer, S.; Klinkenberg, S.

published in
Computer assisted assessment: research into e-assessment
2015

DOI (link to publisher)
10.1007/978-3-319-27704-2_6

document version
Publisher's PDF, also known as Version of record

A Practical Procedure for the Construction and Reliability Analysis of Fixed-Length Tests with Randomly Drawn Test Items

Silvester Draaijer[1(✉)] and Sharon Klinkenberg[2]

[1] Faculty of Psychology and Education, Department of Research and Theory in Education, VU University Amsterdam, Amsterdam, The Netherlands
s.draaijer@vu.nl

[2] Faculty of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, The Netherlands
S.Klinkenberg@uva.nl

Abstract. A procedure to construct valid and fair fixed-length tests with randomly drawn items from an item bank is described. The procedure provides guidelines for the set-up of a typical achievement test with regard to the number of items in the bank and the number of items for each position in a test. Further, a procedure is proposed to calculate the relative difficulty for individual tests and to correct the obtained score for each student based on the mean difficulty for all students and the particular test of a student. Also, two procedures are proposed for the problem to calculate the reliability of tests with randomly drawn items. The procedures use specific interpretations of regularly used methods to calculate Cronbach's alpha and *KR20* and the Spearman-Brown prediction formula. A simulation with R is presented to illustrate the accuracy of the calculation procedures and the effects on pass-fail decisions.

Keywords: Sparse datasets · Classical test theory · Educational measurement · P-value · Reliability

1 Introduction

As the demand for defensibility regarding the quality of online higher education assessment and testing increases, it is crucial that teachers have appropriate tools and guidelines to design and evaluate such tests.

Teachers in higher education can nowadays easily administer formative and summative online achievement tests [1] to student in which test items are randomly drawn from larger item banks. Because items are drawn randomly from an item bank, each student responds to a unique set of test items for the same test. In computer-assisted assessment (CAA) literature, this feature of computer-based testing (CBT) systems is mentioned as a distinctive characteristic of computer-based testing that makes it an attractive alternative to fixed, paper-based tests in view of being able to more systematically address

E. Ras and D. Joosten-ten Brinke (Eds.): CAA 2015, CCIS 571, pp. 47–60, 2015.
DOI: 10.1007/978-3-319-27704-2_6

item quality, prevent item exposure and cheating and provide the possibility of administering tests at multiple instances in time [2].

Teachers have expressed a need to know how many test items should be available in an item bank for test set-ups when such tests are used for formative medium stakes tests or for summative high stakes final examination purposes. In order to respond to that need, it is of importance to first establish the main criteria with which the quality of tests and test items can be judged and, accordingly, how typical set ups of a test and item bank should be designed. As will be suggested, besides content validity, the level of difficulty and reliability of such tests is of main importance.

Further, it is a psychometric challenge to address the issue of difficulty level and reliability with randomly drawn test items and the current CBT systems in use in higher education, such as Questionmark Perception, BTL Surpass, Blackboard, Moodle or Canvas. These systems have limited capabilities for calculating these properties of tests with randomly drawn items. In this paper, this problem will be discussed in more detail and practical procedures for analyzing such tests and estimating their reliability are proposed to optimize fair treatment of students with regards to pass-fail decisions.

First, the case is made to relate the number of test items in an item bank to the number of students taking a test and the number of responses per item required for an analysis with acceptable confidence levels for item and test characteristics. Second, the case is made to systematically adjust individual student scores based on the difficulty level of each individual test. For the latter, statistical procedures to estimate the mean difficulty of a test for students will be described. Finally, estimations for reliability based on classical test theory calculation methods and score adjustment will be presented.

1.1 Background

An important drawback of random item selection from an item bank for each student is that the content validity, reliability and difficulty level of these tests for each individual student are challenging to control. A solution to this problem could be the application of adaptive testing possibilities based on item response theory (IRT). In higher education, however, employing IRT-based approaches is very difficult because it requires advanced technologies and extensive test item development and analysis procedures to develop calibrated test item banks and IRT adaptive tests [3]. Resources and expertise for such applications are in general lacking [4]. Also, the understanding of such procedures by students and the general public is limited, which restricts their acceptability.

In higher education, teachers and support staff resort to better known methods derived from classical test theory (CTT) to assemble and analyze tests and test items. Veldkamp [5] described a procedure for assembling tests of equal difficulty based on CCT when test item banks are available with known values for the difficulty of the test items (p-value) and the correlation values of the test items with the test scores (r_{it}). Veldkamp suggested that item banks should then be structured so that tests could be drawn in a stratified manner to result in tests with equal difficulty and equal deviation. His method built on procedures described by Gibson and Weiner [6]. The main problem with the approach of Veldkamp is that the item characteristics obtained by CTT

inherently are not independent from quality of instruction, quality of circumstances, level and distribution of the test-taker population's ability. This implies that his procedure has fundamental limitations and that an approach is needed that uses obtained item characteristics *after* instruction and administration to students.

2 A Proposal for a Testing Procedure in Higher Education

Teachers in higher education are limited to drawing test items randomly from item banks by the possibilities of the CBT systems at their disposal. These available CBT systems *are* capable of assembling and administering fixed-length tests [7] and of drawing test items randomly without sophisticated drawing algorithms from a pool of test items for each question *position* of a test. This starting point forms a first but feasible step to deploying a construction method that ensures content validity.

2.1 Assumptions

The first assumption for the construction of higher education achievement tests is that there is sufficient content validity: all learning objectives or topics are adequately represented in the test. Further, a rule of thumb in higher education is that summative achievement tests consisting of 4-option multiple-choice questions need at least forty test items of moderate difficulty and acceptable levels of discrimination to reach acceptable levels of measurement precision [8].

A second assumption is that in higher education, the most important decision resulting from an achievement test is whether a student passed or failed the test. For that reason, setting an appropriate cut-off score and ensuring sufficient reliability of a test to minimize false negative or false positive pass-fail decisions is of importance. As every student receives a test with different test items, each student has a test with a unique level of difficulty and reliability. In particular for students who score near the cut-off score, considerations that compensate students with relatively difficult tests are likely to be of importance.

2.2 Proposed Structure of an Item Bank

To ensure a representative selection of test items, a robust procedure is proposed in which, for each *position* in a test, test items are drawn from one specific pool of test items that closely reflects the intended topic and taxonomic cognitive level for that position. Drawing each item from one pool minimizes the chances that test items will be dependent on one another in the test as a whole and will ensure that items will be responded to in a much as possible equally distributed manner. Figure 1 shows this principle of item pool structure and item selection.

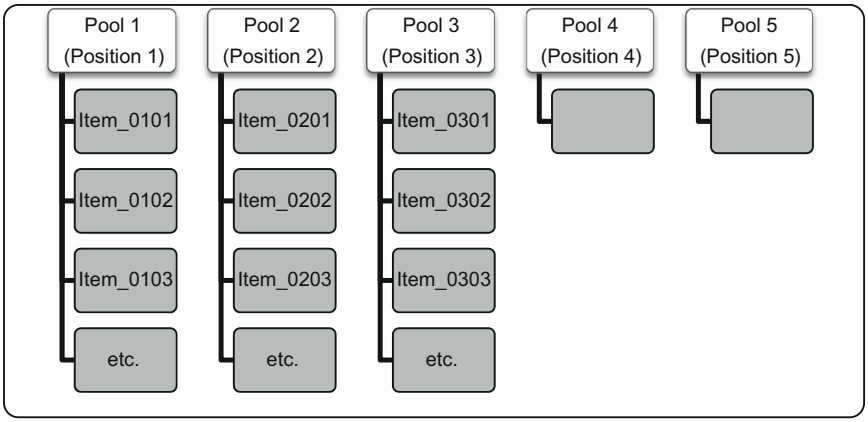

Fig. 1. Example of an item bank structure of an item bank as a reflection of the position of test items in a test

2.3 Number of Test Items for an Item Bank

Though many responses to test items are needed for stable parameter estimations of difficulty and correlation values [9], as a rule of thumb in higher education, 50 responses is regarded as a minimum to be acceptable for decision-making purposes. Taking this minimum as a starting point results in a recommendation for the number of items per position and items according to Eqs. (1) and (2), in which N is the number of students expected to take the exam.

$$n_{(items\ per\ position\ in\ pool)} = \frac{N}{50} \tag{1}$$

$$n_{tot(total\ items\ in\ bank)} = n_{(items\ per\ position\ in\ pool)} * positions\ in\ test \tag{2}$$

2.4 Level of Difficulty for Test Items

It is hard, if not impossible, for teachers in higher education to design test items with known difficulty [10, 11]. Findings from methods and research regarding procedures for item-cloning to control item difficulty are advancing [3, 12], but must be regarded as out of reach for teachers in higher education. Therefore, each student receives test items with different difficulty, resulting in a different level of difficulty for each test. The proposed selection procedure ensures that content validity requirements are met to quite an extent, but does not ensure fairness with regards to difficulty level. The next chapter will address that problem.

3 Estimating the Level of Difficulty for a Test

After construction of an item bank and administration of a test to students, a procedure with the following steps is proposed to estimate the level of difficulty for a test and the level of difficulty for individual students.

First, for each item in the bank, the percentage of students answering the item correctly is calculated. This yields the level of difficulty (proportion correct) p_i for each item. Most CBT systems provide this characteristic for test items and randomly drawn test items by default.

Second, the mean level of difficulty for the test $\bar{p}_{test}$ is calculated by summing the p-value for all items and dividing by the number of test items in the item bank, according to formula (3).

$$\bar{p}_{test} = \frac{\sum_{n_{tot}} p_i}{n_{tot}} \tag{3}$$

Third, according to formula (4), the level of difficulty for the test for each student $\bar{p}_s$ is calculated by summing the p-value for each item a student responded to divided by the number of test items n_s for each student.

$$\bar{p}_s = \frac{\sum_{n_s} p_{i_s}}{n_s} \tag{4}$$

3.1 Correction for Difficulty Levels Between Students

A correction can be made for the level of difficulty for each student in such a way that the level of difficulty of the test will be equal for each student. In the simplest form, this can be done using additive correction. Each student's proportion of correct answers on the test $Prop_{corr}$ will be corrected to $Prop'_{corr}$ as a function of the difference between the mean difficulty of all test items and the mean difficulty for the test of a particular student, as represented in formula (5).

$$Prop'_{corr_s} = Prop_{corr_s} + \left(\bar{p}_{test} - \bar{p}_s\right) \tag{5}$$

After establishing the final adapted score for each student, the cut-off score procedure can be applied. It will be obvious that for a number of students who achieved a score close to the cut-off score, a different decision regarding failing or passing could be made depending on the level of difficulty of their particular test.

A problem with simple additive correction is that students could achieve a proportion correct higher than 100 % if a student scored correct on all items and was provided with a relatively difficult test. In order to overcome this problem, more sophisticated procedures for correction could be applied. For example, correction of scores could be applied only for students with a relatively difficult test and a score close to the cut-off score to prevent false-negative decisions. Or, adjustments could be set so that the amount of

adjustment of the scores runs linearly from zero at the maximum or minimum score to the total corrected score adjustment at the cut-off score. We refer to Livingston [13] for more sophisticated methods for test equating, also incorporating score variance and other considerations.

4 Test Reliability

Well-known methods are available for calculating the reliability of a fixed-length test with a fixed set of test items. The general approach is to calculate Cronbach's alpha α [14] for polytomous items, or $KR20$ (Kuder-Richardson 20 formula) for dichotomous items [15]. In such approaches, variances of item scores and test scores for all students and items are used.

However, in this paper the situation is staged for tests with randomly drawn test items in which the item bank holds more items n_{tot} than are presented to the students N. After administration, the result matrix with the scores for each item for each student is a so-called sparse dataset. The emptiness of these datasets can be in the order of 50 % or more. The large number of empty positions prevents a straightforward calculation of α or $KR20$, in particular because of different interpretations of the number of test items for which calculations need to be carried out and because of calculation procedures in which, for example, list-wise deletion of data occurs. A solution to this problem is to make an estimation of α or $KR20$ using the characteristics of the items in a sparse dataset.

4.1 Lopez-Cronbach's Alpha

The first method for making an estimation of reliability was described by Lopez [16]. In this paper, we refer to this measure as α. The advantage of the Lopez' procedure is that it can be used for both dichotomous and polytomous items. His method uses the correlation matrix of the item scores of items in an item bank. In his approach, the Spearman-Brown prediction formula is conceptualized as in formula (6).

$$\alpha_{tot} = \frac{n_{tot}\bar{r}}{1 + \left(n_{tot} - 1\right)\bar{r}} \tag{6}$$

In (6), $\bar{r}$ is the mean inter-item correlation of the items in the item bank. The procedure that Lopez suggested to calculate α_{tot} is to first calculate the correlation matrix for the items. Second, calculate the mean of the off-diagonal correlation values of the correlation matrix. Third, calculate α using formula (6).

A remaining problem, however, is that the calculated reliability now reflects the situation in which all items in the item bank are used. Based on the assumption of test homogeneity (items have comparable characteristics), a procedure for calculating the mean reliability of all the student's tests is to use the Spearman-Brown prediction formula [17] according to formula (7).

$$\alpha'_{tot} = \frac{k\alpha_{tot}}{1 + (k-1)\,\alpha_{tot}} \tag{7}$$

In formula (7), k is the factor for the relative increase or decrease in the number of items. In the case of items drawn randomly from an item bank, k will always be the proportion of items sampled from the bank divided by the number of items in the bank.

4.2 *KR20*

For dichotomous items, we use a conception of the standard deviation of a test SD_{test} based on Gibson and Weiner [6], using the item-test point-biserial correlation values r_{it_i} of each item i in the item bank and the level of difficulty p_i for each item according to formula (8). The reason for using this formula instead of the regularly used formula for determining SD_{test} is that in formula (8), characteristics of the items distributed to students are sufficient to calculate SD_{test}. Using the SD_{test}, $KR20$ (equal to α for dichotomous items) is calculated according to formula (9).

$$SD_{test} = \sum_{i=1}^{n_{tot}} r_{it_i} \left[p_i \left(1 - p_i\right)\right]^{\frac{1}{2}} \tag{8}$$

$$KR20 = \frac{n_{tot}}{n_{tot} - 1} \left[1 - \frac{\sum_{i=1}^{n_{tot}} p_i \left(1 - p_i\right)}{SD_{test}^2}\right] \tag{9}$$

The values for r_{it_i} and p_i are calculated mostly by default by current CBT systems and could be used to manually calculate $KR20$.

After calculating $KR20$ on the basis of the procedure described above, the Spearman-Brown formula parallel to formula (10) needs to be used again to calculate the mean estimate $KR20'$ for the students based on the number of administered items n_s per student.

$$KR20' = \frac{kKR20}{1 + (k-1)\,KR20} \tag{10}$$

4.3 Test Reliability for Individual Students

When assuming no homogeneity and with dichotomous scoring, the reliability for each *individual* student $KR20_s$ could also be computed by using only the data of the individual items administered to each student n_s, according to formulas (11) and (12).

$$SD_{test_s} = \sum_{i_s=1}^{n_s} r_{it_{i_s}} \left[p_{i_s} \left(1 - p_{i_s}\right)\right]^{1/2} \tag{11}$$

$$KR20_s = \frac{n_s}{n_s - 1} \left[1 - \frac{\sum_{i_s=1}^{n_s} p_{i_s} \left(1 - p_{i_s}\right)}{SD_{test_s}^2}\right] \tag{12}$$

4.4 Simulations for Estimating the Accuracy of Calculated Reliability Parameters

To provide evidence for the degree of accuracy of the procedures described, a simulation was set up using R [18]. In the simulation, two research questions were formulated:

1. To what extent does the correction procedure for the sum scores decrease incorrect pass-fail decisions?
2. How robust are the two presented procedures for calculating Lopez' α and $KR20$ for a typical sparse dataset on the basis of the proposed test construction set up?

In order to determine the robustness of the described procedures, a benchmark for reliability comparison is needed. For this purpose, we ran a simulation where data with known reliability (Cronbach's α) were generated. To achieve this, we sampled data from a multivariate normal distribution from a predefined covariance matrix. Cronbach's α was calculated from this covariance matrix, called sigma (Σ), resulting in a fixed alpha. The covariance matrix had properties that conform to the associated assumptions of homogeneity and equality of variance while also approximating real-world item parameters.

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \dots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \dots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{d1}^2 & \sigma_{d2}^2 & \dots & \sigma_{dn}^2 \end{bmatrix}$$

From this matrix, Cronbach's α was computed using the ratio of mean variance and mean covariance according to formula (13).

$$\alpha = \frac{K\bar{c}}{\bar{v} + (K-1)\bar{c}} \tag{13}$$

Here, K is the number of items, $\bar{v}$ is the average variance for all items (the mean of the diagonal from the covariance matrix Σ), and $\bar{c}$ is the average of all covariances between all items (the off-diagonal from Σ).

By specifying the mean variance and mean covariance, the covariance matrix was used to simulate multivariate data where the underlying α is known. In this example, using $\bar{v} = 1.17$, $\bar{c} = 0.16$ and $K = 400$ results in $\alpha = 0.98$.

We created Σ by sampling the discrimination parameter a for each item from a uniform distribution $a \sim U(0.25, 0.55)$ and applying a residual variance of 1 as is a common assumption within item response theory. Applying this to Σ resulted in:

$$\Sigma = \begin{bmatrix} 1.22 & 0.22 & \dots & 0.26 \\ 0.22 & 1.21 & \dots & 0.25 \\ \vdots & \vdots & \ddots & \vdots \\ 0.26 & 0.25 & \dots & 1.3 \end{bmatrix}$$

Using this covariance matrix, we generated multivariate data $x \sim N_k(\mu, \Sigma)$ consisting of 400 items and 500 students. For later analysis, it was desirable to generate responses based on known abilities θ's and item difficulties β's. We therefore sampled normal θ's $\sim N(0, 1)$ and uniform β's $\sim U(-2, 2)$. Multivariate normal responses were sampled using the mvrnorm() function from the R package MASS written by Ripley et al. [19]. From this, we calculated a response matrix where the binary response was determined by the difficulty, ability, discrimination and covariance structure. We categorized the continuous response by assigning values of 1 when it exceeded the item difficulty β and values of 0 when the continuous response was lower than β. For a detailed description of binary data modeling we refer to De Boeck and Wilson [20].

The following procedure was used for this simulation. We generated a binary response matrix with dimensions of 400 and 500 based on the above method, with a known $\alpha = 0.98$. We calculated Cronbach's α from the response matrix using the cronbach.alpha() function from the ltm package written by Rizopoulos [21], applied the *KR*20 and Lopez' method and then applied Spearman-Brown's formula to all methods. We also calculated $\rho_{xx'}$ by correlating the standardized known student ability θ's with standardized sum scores. This represented the real reliability. From the full response matrix, a sparse matrix was created by randomly sampling 40 responses for every student. The sparse matrix was used to again calculate Cronbach's α, *KR*20, and Lopez and apply the Spearman-Brown correction. In addition to calculating $\rho_{xx'}$ on the sum scores of the sparse matrix, we also calculated the corrected sum scores using the method described in formula (5). This procedure was repeated 10,000 times to get robust estimators and determine their lower and upper bounds based on a 95 % confidence interval. Confidence intervals were calculated using the 2.5 % and 97.5 % quantile scores. The full simulation code can be found in the GitHub repository by Klinkenberg [21].

Furthermore, we calculated the pass-fail rate based on a predefined cut-off score of 60 %. By comparing this to the true pass-fail rate, we created cross tables containing the amount of correct and incorrect decisions.

4.5 Results of the Simulation

The results of the simulation are graphically represented in Fig. 2. The figure shows the calculated reliabilities and the 95 % confidence interval for each method used. The true alpha on which the data were simulated is indicated at the bottom. The Spearman-Brown corrected reliabilities indicate the estimates for 40 sampled items. These should be compared to the lower bound of the correlations between the true θ's and the sum and corrected sum scores in the sparse data. The remainder of the reliabilities, also in the sparse set, estimated the reliability of the full item bank. Note that alpha for sparse data is missing because it could not be calculated with sparse data using R (or other programs), an essential point of this paper.

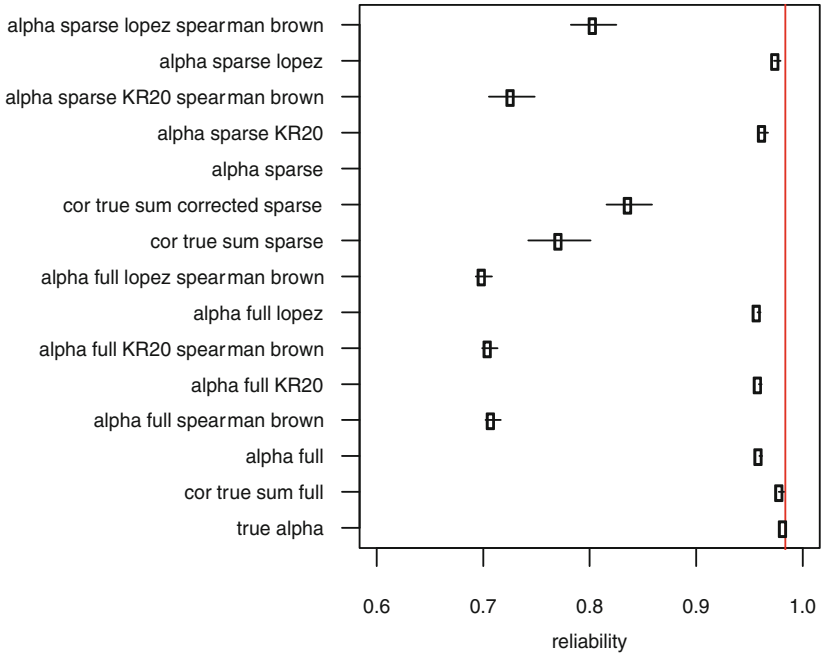

Fig. 2. Reliabilities plotted against true alpha of .98

In Table 1, the correlation between the true abilities and the ability scores $\rho_{xx'}$ shows the true reliability of the test based on a simulated alpha of .98. Further, the table shows the computed values for the different estimation methods using the simulation.

Table 1. Reliabilities as simulated with alpha .98

	Full data		Sparse data		
	Full	Spearman-brown	Sparse	Spearman-brown	Sparse corrected
$\rho_{xx'}$	0.98	.	0.77	.	0.84
CB α	0.96	0.71	.	.	.
*KR*20	0.96	0.71	0.96	0.73	.
Lopez	0.96	0.70	0.98	0.81	.

The table shows that the true reliability of the full dataset corresponds to the true alpha. Also, the true reliability ($\rho_{xx'}$) of the full dataset corresponds to the alpha used to generate the data. This seems a bit strange, as Cronbach's alpha is the lower bound of the true reliability [22]. It would be expected that the alpha used to simulate the data would be lower than the true reliability ($\rho_{xx'}$). We attribute this to the small variations in the estimations due to the large sample size, number of items, the random sampling error

and rounding. The found reliability estimates for the sparse datasets after Spearman-Brown correction for test length show normal values for reliability obtained for achievement test with forty test items (0.77, 0.73, 0.81) and are an indication for the appropriateness of the proposed calculation procedures.

Table 1 further shows that the corrected proportion correct for individual students results in an increase in true reliability compared to the non-corrected sum scores, and a slight increase in true positives. Even when not using correction, the *KR*20 procedure does not result in an overestimation of reliability but in an underestimation (0.73 versus 0.77). Using the Lopez procedure results in an overestimation (0.81 versus 0.77). Further, when applying correction, the Lopez method still yields an underestimation (0.81 versus 0.84).

In Table 2, the sensitivity and specificity of the pass-fail decisions in percentage are given. Of particular interest are the differences in true pass decisions for the sparse and sparse corrected score procedures. This difference is 3 % (from 28 % to 31 %). Though this difference is not large, it has real-world implications; in our simulation, 15 students (3 % of the 500 students) would receive a true-positive instead of a false-negative pass-fail decision, and the number of false-positives would increase by 1 % (9 % instead of 8 %).

Table 2. Sensitivity and specificity of pass-fail decisions

		Full data		Sparse		Sparse corrected	
		Pass	Fail	Pass	Fail	Pass	Fail
True	Pass	37 %	3 %	28 %	12 %	31 %	9 %
	Fail	3 %	57 %	8 %	52 %	9 %	51 %

4.6 Conclusion of the Simulation

In answer to research question 1 regarding the effect of applying a correction procedure for pass-fail decisions, we conclude that correcting the sum scores for mean individual difficulty from sparse data yields a higher reliability (84 % versus 77 %) and lower percentages of false-negative decisions.

With regards to research question 2 concering the robustness of the two presented methods for calculating *KR*20 and Lopez' α, we conclude that the *KR*20 and Lopez methods with Spearman-Brown correction provide practical means for calculating reliability values. However, both methods overestimate the reliability in comparison with the Spearman-Brown correction of the full data matrix. In comparison to the true reliability of the sparse data, we conclude that the *KR*20 method is the most conservative.

5 Conclusion

In this paper, a procedure to construct a fixed length test with randomly drawn items from an item bank has been proposed. The procedure provides guidelines for the set up of a typical test as used in higher education regarding the number of items in the item bank and the number of items for each position in a test. The procedure tries to cater to the need for valid, reliable and fair assessment practices.

Procedures have been proposed for relatively easily obtainable item characteristics to calculate the relative difficulty of individual tests for students and to correct the obtained score for each student based on the mean difficulty of all tests and the difficulty of a particular test.

Two procedures have been presented for solving the problem of calculating the reliability of such tests. This problem needs to be addressed because the test analysis calculation algorithms of current CBT systems used in higher education do not have options for reliability calculation at all or have flawed algorithms for tests with randomly drawn test items. The recommended procedures used a specific interpretation of regularly used methods of calculating α and $KR20$.

The presented simulation showed that the methods described result in valid calculation methods and that the procedure using the $KR20$ approach with Spearman-Brown correction yielded the most conservative estimate.

5.1 Further Research

This study is a first exploration into developing practical means to assess the validity and fairness of achievement tests with randomly drawn test items in higher education using CTT. It answers questions regarding calculation and correction procedures for individual student scores. The study also elicits new research questions.

First, with respect to the estimation procedure of α for sparse data, our study showed different results compared to the original paper by Lopez. In particular, in our simulation, the estimation yielded an overestimation of reliability. Further research is needed to establish why and to what extent these differences occur and are dependent on variables such as number of responses, number of items in the bank and number of items drawn, parameters of student ability difficulty and discrimination distribution of items or use of corrected item-test correlations [23], etc. Obviously, studying the effects of these variables on other estimation methods is needed for further validation of the proposed procedures.

Second, as simulated data were used in our experiment, using real-life data would also provide more insight into the applicability and acceptability of the procedure and calculations.

Third, if tests are provided to students in smaller batches (or even at the level of the individual) running up to the total number of students expected to take the achievement test, methods could be implemented to use streaming calculations. That is, methods could be designed in which item parameters for difficulty and discrimination are set by teachers before test administration and the item parameters could be adjusted as new responses are recorded. The incoming data could then be used to

make better estimations of the item parameters and, hence, better decisions for passing or failing students. This would imply using methods related, for example, to moving averages calculations [24, 25].

5.2 Practical Implications

As our paper has shown, the fairness of pass-fail decisions using randomly drawn test items is hampered because of differences in individual test difficulty. This results in two important implications.

First, when teachers or institutions of higher education design tests in which test items are drawn randomly from an item bank, they should be aware of the differences in individual test difficulty. Although drawing items randomly can be beneficial in view of practical considerations, it has a negative effect on individual students in the false-negative category. Interpreting test results for these tests should be done with caution, and consideration for failed students who encountered more difficult tests is appropriate. Also, attention should be given to evaluating the degree to which teachers and students understand the correction procedure for pass-fail decisions.

Second, a call is made for developers of the CBT software used in higher education to equip their products with features that enable fairer treatment with regard to analysis possibilities and scores correction possibilities when deploying tests with randomly drawn items. Designing such software with a user-friendly interface could be quite a challenge but does not have to be impossible. Our source code is freely available for inspection and further use and development under Creative Commons on Github. This would result in an increased understanding of the characteristics of achievement tests in higher education and in fairer treatment of students.

References

1. Draaijer, S., Warburton, B.: The emergence of large-scale computer assisted summative examination facilities in higher education. In: Kalz, M., Ras, E. (eds.) CAA 2014. CCIS, vol. 439, pp. 28–39. Springer, Heidelberg (2014)
2. Mills, C.N., Potenza, M.T., Fremer, J.J., Ward, W.C.: Computer-Based Testing, Building the Foundation for Future Assessments. Lawrence Erlbaum Associates, London (2002)
3. Glas, C.A.W., Van der Linden, W.J.: Computerized Adaptive Testing With Item Cloning. Appl. Psychol. Meas. **27**, 247–261 (2003)
4. Van Haneghan, J.P.: The impact of technology on assessment and evaluation in higher education. In: Technology Integration in Higher Education: Social and Organizational Aspects, pp. 222–235 (2010)
5. Veldkamp, B.: Het random construeren van toetsen uit een itembank [Random selection of tests from an itembank]. Exam. Tijdschr. Voor Toetspraktijk. **9**, 17–19 (2012)
6. Gibson, W.M., Weiner, J.A.: Generating random parallel test forms using CTT in a computer-based environment. J. Educ. Meas. **35**, 297–310 (1998)
7. Parshall, C.G., Spray, J.A., Kalohn, J.C., Davey, T.: Practical Considerations in Computer-Based Testing. Springer, New York (2002)
8. van Berkel, H., Bax, A.: Toetsen in het Hoger Onderwijs [Testing in Higher Education]. Bohn Stafleu Van Loghum, Houten/Diegem (2006)

9. Schönbrodt, F.D., Perugini, M.: At what sample size do correlations stabilize? J. Res. Personal. **47**, 609–612 (2013)
10. Cizek, G.J., Bunch, M.B.: Standard Setting: a Guide to Establishing and Evaluating Performance Standards on Tests. Sage Publications, Thousand Oaks (2007)
11. Impara, J.C., Plake, B.S.: Teachers' ability to estimate item difficulty: a test of the assumptions in the angoff standard setting method. J. Educ. Meas. **35**, 69–81 (1998)
12. Gierl, M.J., Haladyna, T.M.: Automatic Item Generation: Theory and Practice. Routledge, New York (2012)
13. Livingston, S.A.: Equating Test Scores (without IRT). Educational Testing Service, Princeton (2004)
14. Cronbach, L.J.: Coefficient alpha and the internal structure of tests. Psychometrika **16**, 297–334 (1951)
15. Kuder, G.F., Richardson, M.W.: The theory of the estimation of test reliability. Psychometrika **2**, 151–160 (1937)
16. Lopez, M.: Estimation of Cronbach's alpha for sparse datasets. In: Mann, S., Bridgeman, N. (eds.) Proceedings of the 20th Annual Conference of the National Advisory Committee on Computing Qualifications (NACCQ), pp. 151–155, New Zealand (2007)
17. Spearman, C.: Correlation calculated from faulty data. Br. J. Psychol. 1904–1920 **3**, 271–295 (1910)
18. Team, R.C.: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2015)
19. Ripley, B., Venables, B., Bates, D.M., Hornik, K., Gebhardt, A., Firth, D., Ripley, M.B.: Package "MASS." (2014)
20. De Boeck, P., Wilson, M. (eds.): Explanatory Item Response Models. Springer, New York (2004)
21. Klinkenberg, S.: Simulation for determining test reliability of sparse data sets (2015)
22. Woodhouse, B., Jackson, P.H.: Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: a search procedure to locate the greatest lower bound. Psychometrika **42**, 579–591 (1977)
23. Cureton, E.E.: Corrected item-test correlations. Psychometrika **31**, 93–96 (1966)
24. Lucas, J.M., Saccucci, M.S.: Exponentially weighted moving average control schemes: properties and enhancements. Technometrics **32**, 1–12 (1990)
25. Wei, W.W.: Time Series Analysis. Addison-Wesley, Boston (1994)